Mathematics and Statistics 10(1): 15-24, 2022
DOI: 10.13189/ms.2022.100102 http://www.hrpub.org

Stratification Methods for an Auxiliary Variable Model-Based Allocation under a Superpopulation Model

Bhuwaneshwar Kumar Gupt[1], Mankupar Swer[2,*], Md. Irphan Ahamed[3], B. K. Singh[2], Kh. Herachandra Singh[4]

[1]Department of Statistics, North-Eastern Hill University, Meghalaya, India
[2]Department of Mathematics, North Eastern Regional Institute of Science and Technology, Arunachal Pradesh, India
[3]Department of Mathematics, Umshyrpi College, Meghalaya, India
[4]Department of Mathematics, Manipur University, Manipur, India

Received August 16, 2021; Revised December 9, 2021; Accepted December 23, 2021

Cite This Paper in the following Citation Styles

(a): *[1] Bhuwaneshwar Kumar Gupt, Mankupar Swer, Md. Irphan Ahamed, B. K. Singh, Kh. Herachandra Singh , "Stratification Methods for an Auxiliary Variable Model-Based Allocation under a Superpopulation Model," Mathematics and Statistics, Vol. 10, No. 1, pp. 15 - 24, 2022. DOI: 10.13189/ms.2022.100102.*

(b): *Bhuwaneshwar Kumar Gupt, Mankupar Swer, Md. Irphan Ahamed, B. K. Singh, Kh. Herachandra Singh (2022). Stratification Methods for an Auxiliary Variable Model-Based Allocation under a Superpopulation Model. Mathematics and Statistics, 10(1), 15 - 24. DOI: 10.13189/ms.2022.100102.*

Abstract In this paper, the problem of optimum stratification of heteroscedastic populations in stratified sampling is considered for a known allocation under Simple Random Sampling With and Without Replacement (SRSWR & SRSWOR) design. The known allocation used in the problem is one of the model-based allocations proposed by Gupt [1,2] under a superpopulation model considered by Hanurav [3], Rao [4], and Gupt and Rao [5] which was modified by the author (Gupt [1,2]) to a more general form. The problem of finding optimum boundary points of stratification (OBPS) in stratifying populations considered here is based on an auxiliary variable which is highly correlated with the study variable. Equations giving the OBPS have been derived by minimizing the variance of estimator of the population mean. Since the equations giving OBPS are implicit and difficult for solving, some methods of finding approximately optimum boundary points of stratification (AOBPS) have also been obtained as the solutions of the equations giving OBPS. While deriving equations giving OBPS and methods of finding AOBPS, basic statistical definitions, tools of calculus, analytic functions and tools of algebra are used. While examining the efficiencies of the proposed methods of stratification, they are tested in a few generated populations and a live population. All the proposed methods of stratification are found to be efficient and suitable for practical applications. In this study, although the proposed methods are obtained under a heteroscedastic superpopulation model for level of heteroscedasticity one, the methods have shown robustness in empirical investigation in varied levels of heteroscedastic populations. The stratification methods proposed here are new as they are derived for an allocation, under the superpopulation model, which has never been used earlier by any researcher in the field of construction of strata in stratified sampling. The proposed methods may be a fascinating piece of work for researchers amidst the vigorously progressing theoretical research in the area of stratified sampling. Besides, by virtue of exhibiting high efficiencies in the performance of the methods, the work may provide a practically feasible solution in the planning of socio-economic survey.

Keywords Auxiliary Variable, Estimation Variable, Optimum Boundary Points of Stratification, Superpopulation Model, Stratified Simple Random Sampling with and Without Replacement

1. Introduction

The history of the problem of construction of strata in sampling design dated back to 1950. Dalenius [6] was the

pioneer in this field. Considering the study variable only, he determined the optimum boundary points of stratification (OBPS) for known allocation of sample size to strata. Several other authors including Dalenius and Gurney [7], Aoyama [8], Ekman [9], and Dalenius and Hodges [10] also considered the problem of stratification based on study variable. In practice, the study variable is unknown before survey. This requires one to use another variable, called the auxiliary variable, whose information is available beforehand. The auxiliary variable considered must be highly correlated with the study variable. Problem of finding OBPS and also approximation methods of finding AOBPS for different allocations based on the auxiliary variable which is closely correlated to the study variable was considered by Dalenius and Gurney [7], Taga [11], Singh and Sukhatme [12], Singh [13,14], Singh and Prakash [15], Yadava and Singh [16] and several other authors.

Using a superpopulation model, Singh and Sukhatme [12] obtained the equations giving OBPS and AOBPS based on a concomitant variable for Tschuprow [17] - Neyman [18] optimum allocation (TNOA) and proportional allocation. The problem of allocating the sample size to the strata based on auxiliary variable which is highly correlated with the study variable under a superpopulation model that was first taken into account by Hanurav [3] and Rao [4]. Gupt and Rao [5] considered the same superpopulation model in obtaining allocation of sample size to strata in sampling with probability proportional to size with replacement design. Gupt [1,2] considered a more general form of the same superpopulation model taking into account the existence of intra-stratum correlation coefficient among the units within strata. Under this superpopulation model, Gupt obtained several allocations based on the auxiliary variable for SRSWR and SRSWOR and also proposed some approximations to the allocations obtained. Gupt and Ahamed [19] considered the problem of optimum stratification for a generalized auxiliary variable optimum allocation (GAVPA) obtained by Gupt [1,2]. Gupt and Ahamed [20] and Gupt et al., [21] also worked on stratification for a known allocation obtained by Gupt [1,2] and auxiliary variable optimum allocation proposed by Hanurav [3] respectively under SRSWR & SRSWOR. Ahamed et al., [22] too proposed several methods of stratification for the generalised auxiliary variable optimum allocation (GAVOA) proposed by Gupt [1,2]. On the other hand, Gupt et al., [23] dealt with the problem of construction of strata in cluster sampling with clusters of equal size for allocation proportional to strata total when clusters are considered as sampling units of population and hence proposed several methods of stratification based on auxiliary variable highly correlated with the characteristic under study.

One of the allocations Gupt [1,2] obtained is

$$n_h \propto N_h \sqrt{\bar{X}_h}. \qquad (1)$$

$\forall$ h=1, 2, ..., L and L is the number of strata into which the population N is divided such that $\sum_{h=1}^{L} N_h = N$ under the following superpopulation model :

$$\left.\begin{aligned} \xi(y_i|x_i) &= \alpha + \beta x_i \\ \mathcal{V}(y_i|x_i) &= \sigma^2 x_i \\ \zeta(y_i, y_j|x_i, x_j) &= 0 \end{aligned}\right\} \qquad (2)$$

where α, β, σ^2 are superpopulation parameters with $\sigma^2 > 0$ and ξ, $\mathcal{V}$, ζ denote the conditional expectation, variance and covariance given x_i respectively.

In this paper, we deal with the problem of obtaining the OBPS and AOBPS for the allocation given by (1) under stratified SRSWR design. This paper is divided into six sections. In Section 2, Data and methodology are briefly given. Section 3 contains the extensive derivation for the equations giving OBPS. In Section 4, derivations for methods of finding AOBPS are presented. Empirical illustration of the proposed methods of stratification in several generated populations as well as a live population is given in section 5 along with discussion on their efficiencies. Finally, conclusion is given in Section 6.

2. Data and Methodology

2.1. Data

The data used in the empirical examination of the proposed methods of stratification are three generated populations which follow probability density functions (PDF) viz., Uniform, Exponential and Right Triangular probability density functions, and a live population viz., the number of households and total population of each of 318 villages of Mawshynrut Community and Rural Development Block, West Khasi Hills District of Meghalaya in India based on the population census 2011, taken from 'A Handbook on the Block-Wise Demographic Profile of Meghalaya', published by the Directorate of Economics and Statistics, Meghalaya, India [24].The population per village is taken as the estimation variable y while the number of households per village is taken as the auxiliary variable x.

2.2. Methodology

In the derivation for obtaining methods of stratifications, viz., equations giving OBPS and methods of finding AOBPS, basic statistical definitions, tools of calculus, analytic functions and tools of algebra are used. The AOBPS are derived from OBPS as solutions of the OBPS by using abovementioned techniques of mathematical analysis and some identities developed by Singh and Sukhatme [12] and Ekman [25]. While generating the populations, SPSS software is used in which the relation between study variable y and stratification variable x is obtained subject to the conditions that regression of y on x is linear with slope 45° and σ^2 in $\mathcal{V}(y_i|x_i) = \sigma^2 x_i$ is

determined in such a way that 90% of the total variation of the regression variable *y* is accounted for by the regression. The exponential PDF is truncated in such a way that the area under the curve to the right of the truncation point is 5%. While using approximation methods in stratifying both generated data and live data, numerical integration and differentiation are used wherever required.

In the case of illustrating methods of finding AOBPS in the live data, we need to find the most appropriate PDF followed by x variable of the live data. For this purpose, we use fitdistrplus package in *R*-software to select the best fitted PDF to the x variable of the data. In the determination of σ^2 and g in $\mathcal{V}(y_i|x_i) = \sigma^2 x_i^g$ of the live data, Levenberg-Marquardt algorithm in SPSS is used in the non-linear regression equation $e_i^2 = \sigma^2 X_i^g + u_i$ (Johnston and DiNardo [26]).

3. Derivation of the Equations Giving Optimum Boundary Points of Stratification

The allocation given in (1) can be reduced to

$$n_h = \frac{nW_h\sqrt{\bar{X}_h}}{\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}}\ . \tag{3}$$

Under the superpopulation model (2), Gupt [1,2] obtained the following:

$$\xi\left(\sigma_{y,h}^2|\underline{X}_h\right) = \beta^2\sigma_{x,h}^2 + \sigma^2\bar{X}_h, \tag{4}$$

where $\underline{X}_h' = \left(X_{h1}\ \ X_{h2}\ldots\ X_{hN_h}\right)$ is a vector of *x*-values in the h^{th} stratum.

The variance of estimator of population mean of the stratified SRSWR sampling design for the allocation (3) under the model (2) is given by

$$V(\bar{Y}_{st}) = \left[\sum_{h=1}^{L}\frac{W_h\sqrt{\bar{X}_h}}{n}\right]\left[\sum_{h=1}^{L} W_h\frac{\sigma_{y,h}^2}{\sqrt{\bar{X}_h}}\right]. \tag{5}$$

Taking conditional expectation of (5), we have

$$\xi\left[V(\bar{Y}_{st})|\hat{X}\right] = \left[\sum_{h=1}^{L}\frac{W_h\sqrt{\bar{X}_h}}{n}\right]\left[\sum_{h=1}^{L}\frac{W_h}{\sqrt{\bar{X}_h}}\xi\left(\sigma_{y,h}^2|\underline{X}_h\right)\right], \tag{6}$$

where $\hat{X}=\left(\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_L\right)$.

From (4) and (6), we obtain

$$\xi\left[V(\bar{Y}_{st})|\hat{X}\right] = \frac{\sigma^2}{n}\left[\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right]^2 + \frac{\beta^2}{n}\left[\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right]\left[\sum_{h=1}^{L}\frac{W_h\sigma_{x,h}^2}{\sqrt{\bar{X}_h}}\right] \tag{7}$$

By taking x_h as the point of stratification between h^{th} and $(h+1)^{th}$ strata, we minimize $\xi\left[V(\bar{Y}_{st})|\hat{X}\right]$ by differentiating (7) partially with respect to x_h and then equating to zero as follows:

$$\frac{\sigma^2}{n}\frac{\partial}{\partial x_h}\left[\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right]^2 + \frac{\beta^2}{n}\frac{\partial}{\partial x_h}\left[\left(\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right)\left(\sum_{h=1}^{L}\frac{W_h\sigma_{x,h}^2}{\sqrt{\bar{X}_h}}\right)\right] = 0$$

$$\Rightarrow \frac{2\sigma^2}{n}\left(\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right)\frac{\partial}{\partial x_h}\left(\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right) + \frac{\beta^2}{n}\left[\sum_{h=1}^{L}\frac{W_h\sigma_{x,h}^2}{\sqrt{\bar{X}_h}}\frac{\partial}{\partial x_h}\left(\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right) + \left(\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right)\frac{\partial}{\partial x_h}\sum_{h=1}^{L}\frac{W_h\sigma_{x,h}^2}{\sqrt{\bar{X}_h}}\right] = 0$$

$$\Rightarrow\left[\frac{2\sigma^2}{n}\left(\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right) + \frac{\beta^2}{n}\sum_{h=1}^{L}\frac{W_h\sigma_{x,h}^2}{\sqrt{\bar{X}_h}}\right]\frac{\partial}{\partial x_h}\left(\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right) + \frac{\beta^2}{n}\left(\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right)\frac{\partial}{\partial x_h}\left(\sum_{h=1}^{L}\frac{W_h\sigma_{x,h}^2}{\sqrt{\bar{X}_h}}\right) = 0 \tag{8}$$

While differentiating partially with respect to x_h, all other terms vanish except the h^{th} and $(h+1)^{\text{th}}$ terms, therefore, we have

$$\left[\frac{2\sigma^2}{n}\left(\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right) + \frac{\beta^2}{n}\left(\sum_{h=1}^{L}\frac{W_h\sigma_{x,h}^2}{\sqrt{\bar{X}_h}}\right)\right]\left[\frac{\partial}{\partial x_h}\left(W_h\sqrt{\bar{X}_h}\right) + \frac{\partial}{\partial x_h}\left(W_{h+1}\sqrt{\bar{X}_{h+1}}\right)\right] + \frac{\beta^2}{n}\left(\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right)\left[\frac{\partial}{\partial x_h}\left(\frac{W_h\sigma_{x,h}^2}{\sqrt{\bar{X}_h}}\right) + \frac{\partial}{\partial x_h}\left(\frac{W_{h+1}\sigma_{x,h+1}^2}{\sqrt{\bar{X}_{h+1}}}\right)\right] = 0 \tag{9}$$

If $f(x)$ is the density function of *x*, we have

$W_h = \int_{x_{h-1}}^{x_h} f(x)dx$, and $W_h = \frac{N_h}{N}$, we can get

$$W_h.\frac{\partial\bar{X}_h}{\partial x_h} = (x_h - \bar{X}_h)f(x_h),$$

$$\text{and } \frac{\partial}{\partial x_h}\left(W_h\sigma_{x,h}^2\right) = (x_h - \bar{X}_h)^2 f(x_h).$$

Then, further we can get

$$\frac{\partial}{\partial x_h}\left(W_h\sqrt{\bar{X}_h}\right) = \frac{1}{2\sqrt{\bar{X}_h}}(x_h + \bar{X}_h)f(x_h). \tag{10}$$

Similarly,

$$\frac{\partial}{\partial x_h}\left(W_{h+1}\sqrt{\bar{X}_{h+1}}\right) = -\frac{1}{2\sqrt{\bar{X}_{h+1}}}(x_h + \bar{X}_{h+1})f(x_h) \tag{11}$$

Also,

$$\frac{\partial}{\partial x_h}\left(\frac{W_h\sigma_{x,h}^2}{\sqrt{\bar{X}_h}}\right) = \frac{x_h-\bar{X}_h}{\bar{X}_h\sqrt{\bar{X}_h}}\left[(x_h - \bar{X}_h)\bar{X}_h - \frac{\sigma_{x,h}^2}{2}\right]f(x_h) \tag{12}$$

And

$$\frac{\partial}{\partial x_h}\left(\frac{W_{h+1}\sigma_{x,h+1}^2}{\sqrt{\bar{X}_{h+1}}}\right) = -\frac{x_h-\bar{X}_{h+1}}{\bar{X}_{h+1}\sqrt{\bar{X}_{h+1}}}\left[(x_h - \bar{X}_{h+1})\bar{X}_{h+1} - \frac{\sigma_{x,h+1}^2}{2}\right]f(x_h). \tag{13}$$

Substituting (10), (11), (12) and (13) in (9), we get

$$\frac{1}{\sqrt{\bar{X}_h}}\left[\beta^2\left(\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right)\{2(x_h-\bar{X}_h)^2 - \sigma_{x,h}^2\left(\frac{x_h}{\bar{X}_h}-1\right)\right\} + \{2\sigma^2\left(\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right) + \beta^2\sum_{h=1}^{L}\frac{W_h\sigma_{x,h}^2}{\sqrt{\bar{X}_h}}\right\}(x_h+\bar{X}_h)]$$
$$= \frac{1}{\sqrt{\bar{X}_{h+1}}}[\beta^2\left(\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right)\{2(x_h-\bar{X}_{h+1})^2 - \sigma_{x,h+1}^2\left(\frac{x_h}{\bar{X}_{h+1}}-1\right)\right\} + \left\{2\sigma^2\left(\sum_{h=1}^{L} W_h\sqrt{\bar{X}_h}\right) + \beta^2\left(\sum_{h=1}^{L}\frac{W_h\sigma_{x,h}^2}{\sqrt{\bar{X}_h}}\right)\right\}(x_h+\bar{X}_{h+1})] \quad (14)$$

Equations (14) give the optimum boundary points of stratification for the study variable based on auxiliary variable x.

Hence, we get the following theorem:

Theorem 3.1: For the allocation $n_h \propto N_h\sqrt{\bar{X}_h}$ in stratified SRSWR sampling design under the superpopulation model (2), the optimum boundary points of stratification of the study variable y in terms of auxiliary variable x are given by the equations (14). The equations hold true for SRSWOR design too when the finite population correction is ignored.

4. Derivation of the Equations giving Approximately Optimum Boundary Points of Stratification

Equations (14) can be equivalently written as

$$\frac{1}{\sqrt{\mu_h}}[\beta^2 A\left\{2(x_h-\mu_h)^2 - \sigma_{x,h}^2\left(\frac{x_h}{\mu_h}-1\right)\right\} + \{2\sigma^2 A + \beta^2 B\}(x_h+\mu_h)]$$
$$= \frac{1}{\sqrt{\mu_j}}[\beta^2 A\left\{2(x_h-\mu_j)^2 - \sigma_{x,j}^2\left(\frac{x_h}{\mu_j}-1\right)\right\} + \{2\sigma^2 A + \beta^2 B\}(x_h+\mu_j)]$$
$$\Rightarrow \frac{1}{\sqrt{\mu_h}}[\beta^2 A\left\{2(x_h-\mu_h)^2 - \sigma_{x,h}^2\left(\frac{x_h}{\mu_h}-1\right)\right\} + M(x_h+\mu_h)] = \frac{1}{\sqrt{\mu_j}}[\beta^2 A\left\{2(x_h-\mu_j)^2 - \sigma_{x,j}^2\left(\frac{x_h}{\mu_j}-1\right)\right\} + M(x_h+\mu_j)] \quad (15)$$

where,

$$\left.\begin{aligned}\bar{X}_h &= \mu_h,\\ A &= \left(\sum_{h=1}^{L} W_h\sqrt{\mu_h}\right),\\ B &= \left(\sum_{h=1}^{L}\frac{W_h\sigma_{x,h}^2}{\sqrt{\mu_h}}\right),\end{aligned}\right\} \quad (16)$$
$$M = 2\sigma^2 A + \beta^2 B \text{ and } j = h+1.$$

We observe that finding exact solutions from equations (15) is difficult since the equations are the function of the population parameters which are also functions of the points of stratification; therefore, we find approximate solutions of equations (15). At first, we expand both sides of the equations about the point x_h, which is the boundary point of the h^{th} and j^{th} strata for $j = h+1$ is assumed. Following the approach of Singh and Sukhatme [12] of using Ekman's [25] identity and assuming that the function $f(x)$ possesses various partial derivatives for all x in the range (m, n) with $(n-m) < \infty$, we obtain the series expansion of conditional mean and variance. Initially we assume A and B defined in (16) as constants. Thus, using Taylor's expansion along with the abovementioned techniques, we expand the terms of the Right Hand Side (RHS) of (15) as follows:

$$\mu_j(x) = x_h + \frac{k_j}{2} + \frac{f'(x_h)}{12f(x_h)}{k_j}^2 + \frac{f(x_h)f''(x_h)-f'^2(x_h)}{24f^2(x_h)}{k_j}^3 + O({k_j}^4), \quad (17)$$

$$W_j = f(x_h)k_j + \frac{f'(x_h)}{2}{k_j}^2 + \frac{f''(x_h)}{6}{k_j}^3 + \frac{f'''(x_h)}{24}{k_j}^4 + O({k_j}^5), \quad (18)$$

$$\sigma_{x,j}^2 = \frac{k_j^2}{12} + O\left({k_j}^4\right), \quad (19)$$

where the function f and its derivatives are evaluated at $t = x_h$ in the interval $t \in [x_h, x_{h+1}]$ and $k_j = (x_{h+1}-x_h)$.

Using (17) and (19), the RHS of equations (15) can be expressed as

$$\frac{1}{\sqrt{x_h}}[2Mx_h + \frac{8\beta^2 Ax_h + M}{16x_h}{k_i}^2 + \frac{\{16\beta^2 Ax_h^2 f'(x_h) + 2Mx_h f'(x_h) - 8\beta^2 Ax_h f(x_h) - 3Mf(x_h)\}}{96{x_h}^2 f(x_h)}{k_i}^3 + O({k_i}^4)] \quad (20)$$

Similarly, the LHS of equations (15) can also be expressed as

$$\frac{1}{\sqrt{x_h}}[2Mx_h + \frac{8\beta^2 Ax_h + M}{16x_h}{k_h}^2 - \frac{\{16\beta^2 Ax_h^2 f'(x_h) + 2Mx_h f'(x_h) - 8\beta^2 Ax_h f(x_h) - 3Mf(x_h)\}}{96{x_h}^2 f(x_h)}{k_h}^3 + O({k_h}^4)] \quad (21)$$

where the function f and its derivatives are evaluated at $t = x_h$ in the interval $t \in [x_{h-1}, x_h]$ and $k_h = (x_h - x_{h-1})$.

When the number of strata is large and the strata width k_h is very small, the higher powers of k_h in the expansion can be neglected.

Putting (20) and (21) in equations (15) and taking $g(t) = \frac{8\beta^2 At + M}{t^{3/2}}$, on simplification, we get

$${k_h}^3 g(t)f(t)\left[1 - \frac{\frac{\partial}{\partial x_h}\{g(t)f(t)\}}{g(t)f(t)}\frac{k_h}{2} + O\left({k_h}^2\right)\right] = {k_j}^3 g(t)f(t)\left[1 + \frac{\frac{\partial}{\partial x_h}\{g(t)f(t)\}}{g(t)f(t)}\frac{k_j}{2} + O({k_j}^2)\right] \quad (22)$$

Singh and Sukhatme [12] also obtained the following result by expanding $\sqrt[\lambda]{f(t)}$ about the point $t = y$:

$$\left[\int_y^x \sqrt[\lambda]{f(t)}\,dt\right]^\lambda = k^\lambda f(y)\left[1 + \frac{k}{2}\frac{f'(y)}{f(y)} + O(k^2)\right]$$
$$= k^{\lambda-1}\int_y^x f(t)dt\,[1 + O(k^2)] \quad (23)$$

Using (23) in (22), we obtain,

$$k_h{}^2 \int_{x_{h-1}}^{x_h} g(t)f(t)dt\left[1 + O(k_h{}^2)\right]$$
$$= k_j{}^2 \int_{x_h}^{x_{h+1}} g(t)f(t)dt\left[1 + O(k_j{}^2)\right]$$
$$\Rightarrow k_h{}^2 \int_{x_{h-1}}^{x_h} g(t)f(t)dt = C_1, \text{ a constant.} \quad (24)$$
$$\Rightarrow \int_{x_{h-1}}^{x_h} \sqrt[3]{g(t)f(t)}dt = C_2. \quad (25)$$

Each of the equations (24) and (25) gives the approximately optimum boundary points of stratification as the solutions to equations (15).

But while deriving the relation (24) or (25) we have initially assumed that A and B defined in (16) as constants. Therefore, the relation (24) or (25) can be successfully applied for finding the approximately OBPS provided the values of A and B can be approximately determined.

Expressions for A and B for determining their approximate values:

Following in the line of Singh and Sukhatme [12], and using the expansion

$$\int_{x_h}^{x_{h+1}} \sqrt{t}\,f(t)dt = \left(\sqrt{x_h}f(x_h)\right)k_j + \left(\sqrt{x_h}f(x_h)\right)'\frac{k_j{}^2}{2!}$$
$$+\left(\sqrt{x_h}f(x_h)\right)''\frac{k_j{}^3}{3!} + \left(\sqrt{x_h}f(x_h)\right)'''\frac{k_j{}^4}{4!} + O(k_j{}^5)$$

together with (17) and (18), we get

$$W_j\sqrt{\mu_j} - \int_{x_h}^{x_{h+1}} \sqrt{t}\,f(t)dt$$
$$= \frac{1}{96}\frac{f(x_h)}{x_h^{\frac{3}{2}}}k_j{}^3\left[1 + \frac{\frac{\partial}{\partial x_h}\left\{f(x_h)/x_h^{\frac{3}{2}}\right\}}{f(x_h)/x_h^{\frac{3}{2}}}\frac{k_j}{2} + O(k_j{}^2)\right]$$
$$= \frac{1}{96}k_j{}^2\int_{x_h}^{x_{h+1}}\frac{f(t)}{t^{\frac{3}{2}}}dt\left[1 + O(k_j{}^2)\right]$$
$$= \frac{1}{96}\left[\int_{x_h}^{x_{h+1}} \sqrt[3]{\frac{f(t)}{t^{\frac{3}{2}}}}\,dt\right]^3$$
$$\Rightarrow W_j\sqrt{\mu_j} = \int_{x_h}^{x_{h+1}} \sqrt{t}\,f(t)dt + \frac{1}{96}\left[\int_{x_h}^{x_{h+1}} \sqrt[3]{\frac{f(t)}{t^{\frac{3}{2}}}}\,dt\right]^3$$
$$\Rightarrow \sum_{j=1}^L W_j\sqrt{\mu_j} = \sum_{j=1}^L \int_{x_{j-1}}^{x_j} \sqrt{t}\,f(t)dt$$
$$+\frac{1}{96}\sum_{j=1}^L\left[\int_{x_h}^{x_{h+1}} \sqrt[3]{\frac{f(t)}{t^{\frac{3}{2}}}}\,dt\right]^3 \quad (26)$$

When approximately optimum boundary points of stratification are obtained under equal interval stratification, then

$$\int_{x_h}^{x_{h+1}} \sqrt[3]{\frac{f(t)}{t^{\frac{3}{2}}}}\,dt = \frac{1}{L}\int_m^n \sqrt[3]{\frac{f(t)}{t^{\frac{3}{2}}}}\,dt$$

Therefore, equation (26) becomes

$$A = \sum_{j=1}^L W_j\sqrt{\mu_j}$$
$$= \int_m^n \sqrt{t}\,f(t)dt + \frac{1}{96L^2}\left[\int_m^n \frac{\sqrt[3]{f(t)}}{\sqrt{t}}dt\right]^3, \quad (27)$$

which gives the approximate value of A.

Again, using (17), (18) and (19), we get

$$\frac{W_j\sigma_{x,j}^2}{\sqrt{\mu_j}} = \frac{1}{12}\frac{f(x_h)}{\sqrt{x_h}}k_j{}^3\left[1 + \frac{\frac{\partial}{\partial x_h}\left\{\frac{f(x_h)}{\sqrt{x_h}}\right\}}{\frac{f(x_h)}{\sqrt{x_h}}}\frac{k_j}{2} + O(k_j{}^2)\right]$$
$$= \frac{1}{12}k_j{}^2\int_{x_h}^{x_{h+1}}\frac{f(t)}{\sqrt{t}}dt\left[1 + O(k_j{}^2)\right]$$
$$\Rightarrow \frac{W_j\sigma_{x,j}^2}{\sqrt{\mu_j}} = \frac{1}{12}\left[\int_{x_h}^{x_{h+1}} \sqrt[3]{\frac{f(t)}{\sqrt{t}}}\,dt\right]^3$$
$$\Rightarrow \sum_{j=1}^L\frac{W_j\sigma_{x,j}^2}{\sqrt{\mu_j}} = \frac{1}{12}\sum_{j=1}^L\left[\int_{x_h}^{x_{h+1}} \sqrt[3]{\frac{f(t)}{\sqrt{t}}}\,dt\right]^3 \quad (28)$$

Again, when approximately optimum boundary points of stratification are obtained under equal interval stratification, then we may get

$$\int_{x_h}^{x_{h+1}} \sqrt[3]{\frac{f(t)}{\sqrt{t}}}\,dt = \frac{1}{L}\int_m^n \sqrt[3]{\frac{f(t)}{\sqrt{t}}}\,dt$$

Therefore, equation (28) become

$$B = \sum_{j=1}^L\frac{W_j\sigma_{x,j}^2}{\sqrt{\mu_j}} = \frac{1}{12L^2}\left[\int_m^n \sqrt[3]{\frac{f(t)}{\sqrt{t}}}\,dt\right]^3 \quad (29)$$

The expression (29) gives approximate value of B.

Thus, using (27) and (29), the relations (24) or (25) can be successfully applied for finding the AOBPS. Hence we arrive at the following lemma:

Lemma 4.1: Let the function $f(x)$ be bounded and have various partial derivatives for all values of x in (m, n), then for a given number of strata, under equal interval stratification based on the auxiliary variable, the approximate value of the expressions $\sum_{j=1}^L W_j\sqrt{\mu_j}$ and $\sum_{j=1}^L\frac{W_j\sigma_{x,j}^2}{\sqrt{\mu_j}}$ are given as:

$$\sum_{j=1}^L W_j\sqrt{\mu_j} = \int_m^n \sqrt{t}\,f(t)dt + \frac{1}{96L^2}\left[\int_m^n \frac{\sqrt[3]{f(t)}}{\sqrt{t}}dt\right]^3,$$

and $\sum_{j=1}^L\frac{W_j\sigma_{x,j}^2}{\sqrt{\mu_j}} = \frac{1}{12L^2}\left[\int_m^n \sqrt[3]{\frac{f(t)}{\sqrt{t}}}\,dt\right]^3$.

Taking into consideration the above lemma together with the analytical procedure to the development of methods (24) and (25), we have got the following theorem:

Theorem 4.1: If the function $f(x)$ and $g(x)$ be bounded in (m, n) and various partial derivatives of the functions exist for all values of x in (m, n), for a given number of strata, the approximately optimum boundary points of stratification (AOBPS) on the auxiliary variable x are obtained by taking equal intervals on the cumulative of $\sqrt[3]{g(x)f(x)}$.

5. Numerical Illustrations

5.1. Using Generated Data

For investigating the efficiencies of the proposed equations (14) and (24) or (25) in obtaining the OBPS and AOBPS respectively, the populations generated by following three PDFs are used:

(a) Uniform distribution:

$$f(x) = 1\,, \qquad 1 \le x \le 2$$

(b) Exponential distribution:

$$f(x) = e^{-x+1}\,,$$
$$1 \le x < \infty$$

(c) Right triangular distribution:

$$f(x) = 2(2 - x),$$
$$1 \le x \le 2$$

Proceeding along the line of Gupt and Ahamed [19], we take the regression between y and x to be linear with slope 45°. σ^2 in $\mathcal{V}(y_i|x_i) = \sigma^2 x_i$ is determined in such a way that 90% of the total variation of the regression variable y is accounted for by the regression. The exponential PDF is truncated in such a way that the area under the curve to the right of the truncation point is 5%.

Since it is shown in Section 3 that equations (24) and (25) are equivalent, we use equation (25) in our illustration. To examine the methods of stratification in each of the populations, first we find the solutions of the proposed equations (14) to obtain OBPS and method (25) to obtain approximately OBPS by using successive iterations. For finding the approximately OBPS, we also use numerical integration and differentiation methods. Variance of estimator of population mean at the optimum boundary points of stratification and approximately optimum boundary points of stratification for each of the number of strata L = 2, 3, 4, 5, 6 is calculated. We also calculate the variance of estimator of population mean for equal interval stratification for each of the number of strata L = 2, 3, 4, 5, 6. The relative efficiencies of the proposed methods of stratification (14) and (25) with respect to the equal interval stratification in all the populations generated by the considered PDFs are also separately calculated for each considered number of strata and shown in the following Tables 1, 2 and 3 respectively.

From table 1, we find that the proposed equations (14) work with the same efficiency for the numbers of strata L = 2, 3, 4 and 6 and with slightly higher efficiency for L = 5 when compared with that of equal interval stratification. The approximation method (25) works with slightly less efficiency for L = 2, 3, with the same efficiency for L = 4, 6 and with higher efficiency for L = 5 when compared with that of equal interval stratification. But, for population of uniform PDF, equal interval stratification is considered to be an efficient method of stratification, and from the below results, it is observed that the methods (14) and (25) are almost equally efficient as the equal interval stratification.

Table 1. Uniform Distribution

No. of strata (L)	Equal interval stratification		Stratification by equations (14)		Relative efficiency	Stratification by approximation method (25)		Relative efficiency
	Points	$nV(\bar{Y}_{st})$	Points	$nV(\bar{Y}_{st})$		Points	$nV(\bar{Y}_{st})$	
2	1.5	0.0277	1.5171	0.0278	100	1.4808	0.0283	98
3	1.33333, 1.66667	0.0163	1.3236, 1.6648	0.0163	100	1.3213, 1.6431	0.0168	97
4	1.25, 1.5, 1.75	0.0128	1.2567, 1.4790, 1.7341	0.0128	100	1.2428, 1.4808, 1.7254	0.0128	100
5	1.2, 1.4, 1.6, 1.8	0.0126	1.2128, 1.3697, 1.5302, 1.7450	0.0121	104	1.1960, 1.3847, 1.5779, 1.7750	0.0119	106
6	1.16667, 1.33333, 1.50000, 1.66667, 1.83333	0.0110	1.1704, 1.3188, 1.5051, 1.7175, 1.8729	0.0110	100	1.1650, 1.3213, 1.4808, 1.6432, 1.8083	0.0110	100

Table 2. Exponential Distribution

No. of strata (L)	Equal interval stratification		Stratification by equations (14)		Relative efficiency	Stratification by approximation method (25)		Relative efficiency
	Points	$nV(\bar{Y}_{st})$	Points	$nV(\bar{Y}_{st})$		Points	$nV(\bar{Y}_{st})$	
2	2.5	0.1983	2.1103	0.1751	113	2.0201	0.1747	114
3	2.0, 3.0	0.1210	1.6968, 2.5962	0.1037	117	1.6269, 2.4913	0.1064	114
4	1.75, 2.50, 3.25	0.0888	1.6521, 2.3273, 3.1638	0.0857	104	1.4528, 2.0205, 2.7653	0.0785	113
5	1.6, 2.2, 2.8, 3.4	0.0807	1.3817, 1.8651, 2.4747, 3.2262	0.0691	117	1.3545, 1.7766, 2.2922, 2.9451	0.0673	120
6	1.5, 2.0, 2.5, 3.0, 3.5	0.0708	1.3293, 1.6927, 2.0493, 2.5344, 3.2262	0.0568	125	1.2913, 1.6271, 2.0206, 2.4915, 3.0724	0.0556	127

Table 3. Right Triangular Distribution

No. of strata (L)	Equal interval stratification		Stratification by equations (14)		Relative efficiency	Stratification by approximation method (25)		Relative efficiency
	Points	$nV(\bar{Y}_{st})$	Points	$nV(\bar{Y}_{st})$		Points	$nV(\bar{Y}_{st})$	
2	1.5	0.0252	1.4019	0.0222	114	1.3846	0.0223	113
3	1.33333, 1.66667	0.0158	1.2486, 1.5373	0.0132	120	1.2484, 1.5346	0.0132	120
4	1.25, 1.5, 1.75	0.0117	1.1921, 1.4129, 1.6577	0.0101	116	1.1842, 1.3847, 1.6169	0.0100	117
5	1.2, 1.4, 1.6, 1.8	0.0097	1.1905, 1.3880, 1.5748, 1.7566	0.0089	109	1.1468, 1.3016, 1.4726, 1.6696	0.0089	109
6	1.16667, 1.33333, 1.5, 1.66667, 1.83333	0.0087	1.1665, 1.3340, 1.4981, 1.6603, 1.8071	0.0087	100	1.1222, 1.2484, 1.3847, 1.5346, 1.7064	0.0080	109

The high relative efficiency of the proposed equations (14) as well as the approximation method (25) in the population of exponential PDF, given in table 2, show that both the methods (14) and (25) perform with much higher efficiency as compared to the equal interval stratification. It is also found that the approximation method (25) works with more or less the same efficiency as the proposed equations (14).

From table 3, we find that in the population of right triangular PDF, the proposed equations (14) work with much higher efficiency for L = 2, 3, 4, 5 and with same efficiency for L= 6 when compared with that of equal interval stratification. The approximation method (25) perform with much higher efficiency for L = 2, 3, 4, 5, 6 when compared with that of equal interval stratification. In this case too, the approximation method (25) performs with almost the same efficiency as the proposed equations (14).

5.2. Using Live Data

We also examine the performances of the equations (14) and the approximation method (25) in obtaining optimum points of stratification in the live population. The live population data we consider in our illustration are the number of households and total population of 318 villages of Mawshynrut Community and Rural Development Block, West Khasi Hills District of Meghalaya in India based on the population census 2011, taken from 'A Handbook on the Block-Wise Demographic Profile of Meghalaya', published by the Directorate of Economics and Statistics, Meghalaya, India [24]. The population per village is taken as the estimation variable y while the number of households per village is taken as the auxiliary variable x.

The methods of stratification (14) and (25) are illustrated in this population too in the same way that we have conducted in generated populations; the points of stratification, variance of estimator of population mean and

efficiencies with respect to equal interval stratification are calculated for each considered number of strata and shown in Table 4.

For using the approximation method (25) to find the AOBPS, we determine the most suitable PDF that the auxiliary variable x follows and estimate some stipulated parameters of the population as follows:

On applying linear least square regression technique between the response variable y and explanatory variable$x$,we get coefficient of determination R^2=0.98244, intercept, $\alpha = -3.841$, β = 5.901. It shows the auxiliary variable x is highly correlated with study variable y.

For estimating the parameters σ^2 and g, the non-linear regression equation $e_i^2 = \sigma^2 X_i^g + u_i$ (Johnston and DiNardo [26]) is used.

On using the Levenberg-Marquardt algorithm in SPSS, we obtain the estimates $\sigma^2 = 4.256$ and $g = 1.33$, it shows the population conforms to heteroscedastic regression superpopulation model of level of heteroscedasticity 1.33.

For determining the suitable density function that x follows, we divide each value of the variable x by 100 and then using the fitdistrplus package in *R*-software, we fit a number of known PDFs in the data obtained from dividing values of x variable by 100. Using the methods - Maximum Likelihood Estimation (MLE), Moment Matching Estimation (MME) and Quantile Matching Estimation (QME) one after another, we find the most suitable PDF that the x variable of the live data follows by comparing the values of LL (log likelihood), AIC (Akaike Information Criteria), BIC (Bayesian Information Criteria) and standard errors $(s.e)$ obtained when fitted to various PDFs. Among all the distribution functions tried, we find the log-normal PDF to fit best to the said data.

Thus, the log-normal PDF $f(x)$ followed by the x variable is:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right],\ x > 0 \qquad (30)$$

where the population is characterized by the following parameters estimated by MLE method:

$$Mean\ log,\ \mu = -1.18469,\ s.d\ log,\ \sigma = 0.77042,$$
$$s.e(\mu) = 0.04320,\ s.e(\sigma) = 0.03054,$$
$$LL = 8.44963,\ AIC = -12.89926,$$
$$BIC = -5.37515.$$

Using the PDF (30) along with the estimates of parameters, we examine the performance of the approximation method (25) in stratifying the live population into *L*=2, 3, 4, 5, 6 and the efficiency is compared with respect to equal interval stratification and shown in Table 4.

From table 4, we find that the proposed equations (14) as well as the approximation method (25) work with strikingly high efficiency for $L = 2, 3, 4, 5$ and 6 when compared with that of equal interval stratification. It is also seen that the equations (14) perform better than the approximation method (25) for $L = 2, 3$ and 4. However for $L = 5$ and 6, the approximation method (25) performs better than the equations (14). Overall, we find that the equations (14) as well as the approximation method (25) perform excellently in stratifying the live population.

Table 4. Live Data

No. of strata (L)	Equal interval stratification		Stratification by equations (14)		Relative efficiency	Stratification by approximate method (25)		Relative efficiency
	Points	$nV(\bar{Y}_{st})$	Points	$nV(\bar{Y}_{st})$		Points	$nV(\bar{Y}_{st})$	
2	118	21134.48	60.34	11855.87	178	51.43	12455.60	170
3	79.33, 156.67	12202.61	37.26, 88.22	6193.02	197	31.27, 82.08	6240.47	195
4	60, 118, 176	8252.20	34.20, 72.37, 129.27	4435.23	186	23.31, 51.91, 103.72	4459.06	185
5	48.4, 94.8, 141.2, 187.4	5491.06	29.11, 57.40, 99.58, 165.75	3086.75	178	18.98, 38.82, 68.57, 119.87	2841.99	193
6	40.67, 79.34, 118.01, 156.68, 195.35	4154.56	23.08, 43.14, 73.95, 118.67, 186.88	2616.75	159	16.23, 31.45, 52.01, 82.34, 132.41	2231.44	186

6. Conclusion

We find that in the populations generated by the uniform, exponential and right triangular PDFs, the proposed equations (14) and the approximation method (25) for obtaining optimum boundary points of stratification and approximately optimum boundary points of stratification respectively perform with high efficiency for all numbers of strata L when compared with that of equal interval stratification. We also find that the proposed equations (14) and approximation method (25) for giving OBPS and approximately OBPS respectively perform with almost same efficiency in all the populations generated by the considered density functions. Similarly, for a randomly chosen live population, the proposed equations (14) and the approximation method (25) perform with exceptionally high efficiency for all numbers of strata *L* when compared with that of equal interval stratification. We also find that both the methods perform with almost same efficiency in stratifying the live population. Based on the above observations, it is therefore justified that both the methods (14) and (25) can be effectively applied to stratify heteroscedastic population of varied levels of heteroscedastic levels besides considered level of heteroscedasticity one in the superpopulation model (2). The methods proposed here have shown robustness in their applicability to some extent. This research work may contribute an interesting share to the ongoing theoretical research in stratified sampling as well as give practically viable solution in survey planning. Although the methods are obtained under SRSWR design, the results hold true for SRSWOR design too when finite population correction is ignored.

REFERENCES

[1] B.K. Gupt. Sample size allocation for stratified sampling under a correlated superpopulation model. METRON-International Journal of Statistics, Vol. LXI, No. 1, 35-52, 2003.

[2] B.K. Gupt. Allocation of sample size in stratified sampling under superpopulation models. LAP Lambert Academic Publishing, Saarbrucken, Deutschland/Germany, 2012.

[3] T.V. Hanurav. Optimum Sampling Strategies and Some Related Problems, unpublished Ph.D thesis submitted to the Indian Statistical Institute, 1965.

[4] T.J. Rao. On the allocation of sample size in stratified sampling. Ann. Inst. Stat. Math, Vol. 20, 159-166, 1968.

[5] B.K. Gupt, T.J. Rao. Stratified PPS sampling and allocation of Sample size. Journal of the Indian Society of Agricultural Statistics, Vol. 50(2), 199-208, 1997.

[6] T. Dalenius. The Problem of optimum stratification-I. Scandinavian Actuarial Journal, Vol. 33, 203-213, 1950.

[7] T. Dalenius, M. Gurney. The problem of optimum stratification II. Scandinavian Actuarial Journal, Vol. 34, 133-148, 1951.

[8] H. Aoyama. A study of stratified random sampling. Annals of the Institute of Statistical Mathematics, Vol. 6, 1-36, 1954.

[9] G. Ekman. An approximation useful in univariate stratification. The Annals of Mathematical Statistics, Vol. 30, 219-229, 1959.

[10] T. Dalenius, J.L. Hodge. Minimum variance stratification. Journal of the American Statistical Association, Vol. 54, 88-101, 1959.

[11] Y. Taga. On optimum stratification for the objective variable based on concomittant variable. Annals of the Institute of Statistical Mathematics, Vol. 19, 101-130, 1967.

[12] R. Singh, B.V. Sukhatme. Optimum stratification. Annals of the Institute of Statistical Mathematics, Vol. 21, 515-528, 1969.

[13] R. Singh. An alternate method of stratification on the auxiliary variable. Sankhya, Vol. 37, 100-108, 1975.

[14] R. Singh. On optimum stratification for proportional allocation. Sankhya, Vol. 37, Pt. I, 109-115, 1975.

[15] R. Singh, D. Prakash. Optimum stratification for equal allocation. Annals of the Institute of Statistical Mathematics, Vol. 27, 273-280, 1975.

[16] S.S. Yadava, R. Singh. Optimum stratification for allocation proportional to strata totals for simple random sampling scheme. Communications in Statistics: Theory and Methods, Vol. 13, No. 22, 2793-2806, 1984.

[17] A.A. Tschuprow. On mathematical expectation of the moments of frequency distributions in the case of correlated observations. Metron, Vol. 2, 461-493, 1923.

[18] J. Neyman. On two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society, Vol. 97, No. 4, 558-625, 1934.

[19] B.K. Gupt, M.I. Ahamed. Optimum stratification for a generalized auxiliary variable proportional allocation under a superpopulation model. Communications in Statistics-Theory and Methods, Published online: 25 July 2020. https://doi.org/10.1080/03610926.2020.1793203

[20] B.K. Gupt, M.I. Ahamed. Construction of strata for a model-based allocation under a superpopulation model, Journal of Statistical Theory and Applications, Vol. 20, No. 1, 46-60, 2021.

[21] B.K. Gupt, M.I. Ahamed, M. Phukon. Optimum stratification for an auxiliary variable optimum allocation under a superpopulation model, Advances and Applications in Statistics, Vol. 67, No. 1, 1-20, 2021.

[22] M. I. Ahamed, B. K. Gupt and M. Phukon. Methods of stratification for a generalised auxiliary variable optimum allocation, Mathematics and Statistics, Vol. 9(5), 617-629, 2021. DOI: 10.13189/ms.2021.090501.

[23] B. K. Gupt, F. Lalthlamuanpuii and M. I. Ahamed. Choice of strata boundaries for allocation proportional to stratum cluster totals in stratified cluster sampling. Mathematics and Statistics, Vol. 9(5), 697-710, 2021. DOI: 10.13189/ms.2021.090509.

[24] Directorate of Economics and Statistic Meghalaya. A Handbook on the Block-Wise Demographic Profile of Meghalaya – 2018. Government of Meghalaya, Shillong, 2018. http://megplanning.gov.in/handbook.htm

[25] G. Ekman. Approximate expressions for the conditional mean and variance over small intervals of a continuous distribution. The Annals of Mathematical Statistics, Vol. 30, No. 4, 1131-1134, 1959.

[26] J. Johnson, J. DiNardo. Econometric methods. 4th edition. Singapore: McGraw Hill, 2007.